




Rheinner Guide

.....



RAID Storage



Rheinner
Document
Imaging & Workflow
Technology
Series

techinfocenter.com

[click here](#)

RAID STORAGE

Written and Produced by The Rheininner Group

Distributed by techinfocenter.com (*click here*)

TABLE OF CONTENTS

Introduction	2
Definition of RAID	2
Advantages of RAID	4
The Importance of RAID for Imaging.....	7
RAID Functionality	9
RAID Performance Issues	14
RAID in an Imaging Environment	15
Author Biography	19
Copyright Information	19

The Rheininner Group

The Rheininner Group is a leading research, consulting and education firm in the document imaging, management and workflow industry. Its Certified Document Imaging Architect (CDIA) Education Program, which covers many of the same issues addressed by The Rheininner Group's Technology Guides, is the most popular training program in the imaging industry. For more information on The Rheininner Group, CDIA course schedules, or to obtain help designing and implementing document imaging and workflow systems, please call 781-741-8100 or visit our web site, at www.rheininner.com.

Copyright© 1996, 1997, 1998, 1999 The Rheininner Group, 50 Derby Street, Bare Cove Executive Park, Hingham, MA 02043, Tel:(781) 741-8100 Fax:(781) 741-5885, e-mail: rwasner@rheininner.com. This guide along with many others are available on www.techinfocenter.com, a free online resource dedicated to document systems technology.

The Rheininner Group is a proud sponsor of the Certified Document Imaging Architect™ (CDIA) certification program.

Special thanks to: Ben Sigby for the cool cover, and Holliday Versoy Langille for cool layout and publishing.

INTRODUCTION

This guide presents an introduction to RAID technology and the particular benefits of utilizing RAID in an imaging environment. The guide describes the concepts and issues behind the technology, provides insight into its deployment in document imaging environments, and defines the common terminology associated with the technology.

DEFINITION OF RAID

RAID stands for Redundant Array of Independent or Inexpensive Drives. The original specification for RAID was developed by engineers at the University of California at Berkeley in the late 1980s. The specification was built on the premise that the cost of mass storage could be reduced by combining large numbers of small, inexpensive disks into a single volume to take the place of larger, expensive disks.

In addition, the design goals included providing greater fault tolerance for disk drives, and the ability to protect access to disk data in the event of disk drive failure. Obviously inexpensive disks do not necessarily provide for the greatest level of reliability. For this reason RAID is currently referred to as Redundant Array of Independent Disk Drivss.

Basic Terminology and Concepts: Disk Drive Functionality

In order to understand some of the technology behind RAID it is necessary to explain the basic concepts and technologies behind magnetic disk drives. This section is designed for people who have limited technical knowledge of hard disk drive functionality.

The way in which a disk drive works is not unlike a record player. You place the media on the turntable, move the stylus over the top of the record and place it on the first set of grooves at the outermost edge of the record; the machine plays whatever is contained on those first few tracks. It then proceeds to play the entire record, unless interrupted.

A hard disk drive works in the same way except that the stylus is called the drive head, and the vinyl media is called the platter. Instead of having one read head, as compared to the record player stylus, disk drives have multiple heads. While record players can only play back information, disk drives can both record and playback.

On a vinyl record, or even a CD for that matter, the information is recorded in a series of sequential grooves starting at the outermost edge. The grooves continue on in a spiral until the center of the record where the spiral ends. Disk drives work in a similar way except that information is stored in concentric rings rather than one single ongoing spiral. Each circle is called a track. The drive media rotates inside the drive case at a constant velocity. The drive head can move along the various tracks and briefly touch the surface of any track to either write or playback whatever is located there. It must, however, remember the specific location of that touchdown along the track. In order to coordinate that location, the entire media surface is organized into tracks and sectors (sectors can be visualized as slices of a pie.)

All disks have to be formatted in order to be ready to receive data. Formatting initializes the disk by logging an index of all available tracks and sectors on the media surface. When the disk drive head receives an instruction to place a piece of data on an available portion of the disk media, it checks a reference kept by the computer to make sure that it puts the information into an available sector coordinate. The chunk of information it places onto the surface is called a block. When writing a single file to the disk, blocks of data belonging to that file are scattered onto any available locations and an index of all blocks belonging to this file is kept by the computer. When a file is deleted the computer simply erases the index, and makes those sector locations available for the next write.

As drive capacities have increased, the drive heads have been made smaller and smaller. In this way much more information can be placed on, and accessed from, the same size surface. This means that track and sector locations must be accurate within very small variances each time the head is sent to a location. In order to ensure such precision the mechanical workings of the drive are placed under the control of a semiconductor “brain” called the drive controller. The drive controller manages all the functions related to the drive and manages all calibration.

The performance of all storage devices is limited by their ability to receive data from a computing system and their ability to deliver data to

the computing system, a process referred to as Input and Output (I/O). In general terms, the performance of drives is measured by many things, including access speed and data transfer rate. The access speed is the measure, in milliseconds, of how quickly on average the drive can locate a particular piece of data on the disk surface. After the data is located, the second speed measurement of importance is how fast the information can be delivered to the computer for processing. This is called the data transfer rate, which is measured in megabytes per second. There are important differences between occasional data transfer rates and sustained transfer rates. The sustained transfer rate means that the drive can transfer the amount of data to and from the drive on an on-going basis as opposed to a one-time transfer.

The data transfer rate is also dependent on the nature of the interface between the disk drive and the computer. The most common computing interface is called SCSI (Small Computer System Interface). SCSI is often further defined in terms of SCSI and SCSI 2. SCSI 2 includes two specifications: one for Fast SCSI that allows for greater data transfer rates, and one for Wide SCSI, which enables the transfer of larger chunks of data. Data transfer rates increase dramatically for SCSI 2. Depending on the mechanism, single drives are rarely capable of transferring more than 4 megabytes (MB) per second on a sustained basis. Yet SCSI 2 is capable of achieving data transfer rates in excess of 20MB per second. In order to achieve this type of output, data must be split across several disk drive mechanisms. Enter RAID technology!

ADVANTAGES OF RAID

In principle, stacking a handful of disk drives together should provide numerous advantages based on the principle that the whole will be greater than the sum of its parts. Each disk drive mechanism is a self-contained functional storage mechanism containing heads and media. In the aggregate, all those heads, all that media, and all those controllers should provide a greater performance level than each individual component. By combining disk drives, configuring them differently, and placing them under a separate controller managed with separate software, RAID technology does in fact provide performance greater than the sum of its parts.

Performance

By placing multiple drives in an array (an orderly arrangement of drives), and accessing all of them simultaneously, performance can be greatly enhanced. Instead of being limited by the data transfer rate of each individual component, in a RAID array 20MB per second or more can be achieved when using a SCSI 2 interface. This is five times faster than the performance that could be delivered by the fastest single disk drive on the market.

Capacity is also increased. If each drive in the array is a 1 gigabyte (GB) drive, and there are five drives, all of the storage capacity can be combined so the RAID array provides a 5GB volume. To the computer it looks like a single hard disk capable of storing 5GB. RAID arrays can be configured to accept terabytes (TB) of data. A terabyte is 1,000 gigabytes and although it is an enormous amount of data, volumes like these are not entirely uncommon in the realm of document imaging. To put this in perspective, the capacity of a floppy disk is roughly 1,000 times smaller than a single gigabyte.

Protection Against Failed Disks

A second design criteria for the RAID specifications was the ability to ensure the integrity of the storage subsystem in the event a single drive experienced failure. In normal processing environments, if the disk drive fails during a particular access then all of the data on the drive becomes unavailable and can no longer be accessed. Drive failure is different from a bad file on the disk. A bad file means that the data in that particular file has been corrupted and is not readable, but the rest of the disk may be fine.

File errors can be caused by malfunctioning disks, malfunctioning applications, or a malfunctioning operating system. There is little that can be done to guard against these last two types of failures. As far as the disk drive mechanism is concerned, it performed as it was supposed to. The drive cannot discern between good and bad data so it continues to write whatever is sent to it. However, if the bad files were caused by a malfunctioning device within the RAID array, then the information destined for that drive can be reconstructed. Within the RAID specification, data loss is prevented by providing redundancy in one of two ways:

- 1) **Mirroring:** Mirroring means that whatever is written to one device is also simultaneously written to another. This is an expensive proposition because it requires the purchase of two hard disks, each of which stores the same data. Mirroring is an effective way to protect against failure because it is highly unlikely that both drives will fail at the exact same time. Mirroring is basically data duplication and it performs the same function as creation of a data backup. The only difference is that both the backup and the original write are simultaneously performed and therefore are redundant.
- 2) **Parity:** Parity checking maintains a CHECKSUM of how data is written across a number of disk drives. If one of the pieces of data is accidentally lost due to a drive error, the RAID parity scheme can reconstruct the missing piece. This works as follows:

Assume you have a column of numbers:

$$\begin{array}{r}
 3 \\
 5 \\
 6 \\
 7 \\
 9 \\
 \hline
 30
 \end{array}$$

Adding all the numbers up, the result is 30. Let's suppose that one of the numbers gets erased.

3	3
x	5
6	6
7	7
9	9
25	30
25 - 30 = (5)	

Because you know that the checksum was thirty, you can calculate that the missing number must have been 5. As long as no more than one number is erased, this method will always allow you to recover the missing piece.

In a RAID array set up to use parity, no two values from a group are placed on the same disk drive. If any single drive goes down, the missing

component can always be reconstructed. While this was a mathematical example, in actual use the RAID system utilizes an algorithm called “EXCLUSIVE-OR-CHECKSUM”, which is more applicable to binary data.

While failure of disk drives is one dimension of fault tolerance, guarding against the failure of the disk drive controller is another important feature of RAID. This feature is implemented in some RAID architectures. When a controller fails, the secondary or backup controller immediately assumes the responsibility of the failed device to continue operation.

Battery backup of RAID cache is another feature implemented in some RAID architectures where built-in facilities keep the storage system running long enough to complete a processing cycle, before shutting down.

Flexibility

By providing dramatic increases in data transfer rates and providing data redundancy as well, RAID systems are highly flexible and configurable storage subsystems. Of course the flexibility comes at a price. Both full redundancy and high performance data transfer can not be accomplished at the same time. The ability to perform either should be thought of on a sliding scale with 100% redundancy of data at one end and 100% ability to access all disk drives simultaneously at the greatest speed, at the other end. The RAID specifications use a system of RAID levels, described in a later section, to specify the range of support for each. Some RAID architectures permit more than one level to operate in the same array, and to expand the capacity of the array with varying capacity disk drives, as well as provide different levels of RAID per user.

THE IMPORTANCE OF RAID FOR IMAGING

RAID provides an imaging environment with many benefits. Because document imaging is extremely I/O-intensive, many of the storage issues in imaging appear to be custom designed for the RAID environment. The constant movement of files places extraordinary demands on the storage subsystem requiring high speed

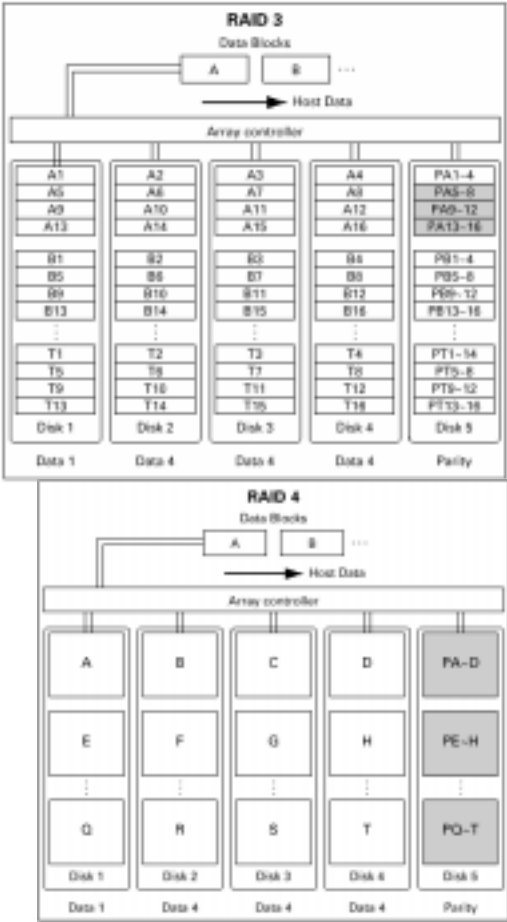
performance, fault tolerance, and high capacity. Some of the storage needs for document imaging systems are as follows:

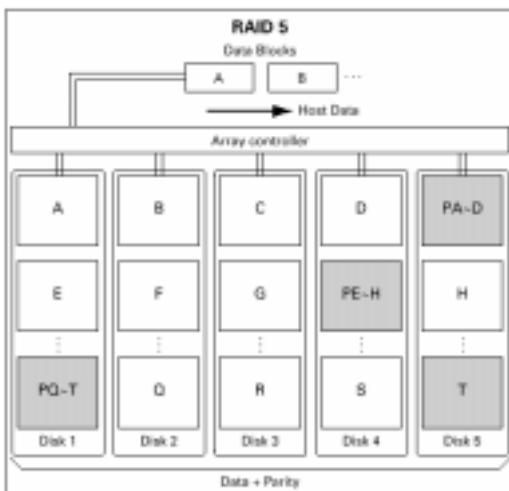
- **Storage of large quantities of information on-line:** In an imaging environment, capturing 10,000 pages per day will utilize 2.5GB of storage for just one week's worth of document input. If the application requires that the documents be on-line and immediately available, then the monthly storage, just for new arrivals, will be in excess of 10GB. With 100 operators processing 100 documents per hour each day, an additional 4GB of images must be accessed from a storage device during each work day. If all documents are kept on-line for the weekly processing period, then 20GB or more of image data may need to be kept on-line for rapid access.
- **Uninterrupted access to high speed storage:** Depending on system volume and storage subsystem configuration, retrieval times may be dramatically extended and can slow down the entire process if every user is going to the same optical disk jukebox for the work he or she needs. Because optical jukeboxes rely on the robotic arm to move disks from storage into the disk drive, multiple simultaneous requests can cause the disk swapping to reach frenetic proportions. Optical jukeboxes can not respond efficiently in such situations. For this reason active documents are best kept on high speed magnetic media. The media of course must be resistant to failure. If any one component fails, the entire work process fails.
- **Flexible capacity for high-speed storage systems:** High-speed storage needs are frequently unpredictable, particularly during peak periods. The ability to add capacity to a single device, or to add additional devices, is a significant benefit of RAID in an imaging environment.
- **Pre-fetching and Caching:** It is common in imaging environments to either pre-fetch or cache data from slower storage devices utilized for long term and permanent storage. Pre-fetching simply means that the anticipated workload is transferred from a jukebox to high-speed magnetic storage during off hours, so that it is available before a request for that information is made. Caching means to store frequently-accessed information temporarily on a high-speed storage device. A third way to improve system response time is to buffer slower devices by holding work on a steady supply source such as a high-speed RAID array in order to make it available for both users and slower storage devices.

RAID FUNCTIONALITY

While RAID has a standard set of specifications, there are a number of ways of implementing the array of disks. One of the main effects of the RAID specifications is to provide multiple approaches to modifying the I/O activities of a series of drives bound together as one complete volume. Thus the specifications are a series of tradeoffs between input performance and output performance. The increase in performance comes at the expense of redundancy. In other words, if you want redundant data you can't have maximum input and output performance, and vice versa.

Disk Array: RAID Levels 3-5





Levels of RAID

There are six levels of RAID: RAID 0 through RAID 5 (which is actually the sixth level.) There are also additional levels of RAID, such as RAID 7, which are extensions to the original specifications and are, in the minds of some experts at least, questionable enhancements to the first six levels.

Not all RAID systems support all of the features inherent in the RAID levels, and some exceed them. Secondly not all RAID devices support all six RAID levels because not all RAID levels are simultaneously applicable. Each RAID level has both a colloquial definition along with the official RAID level designation.

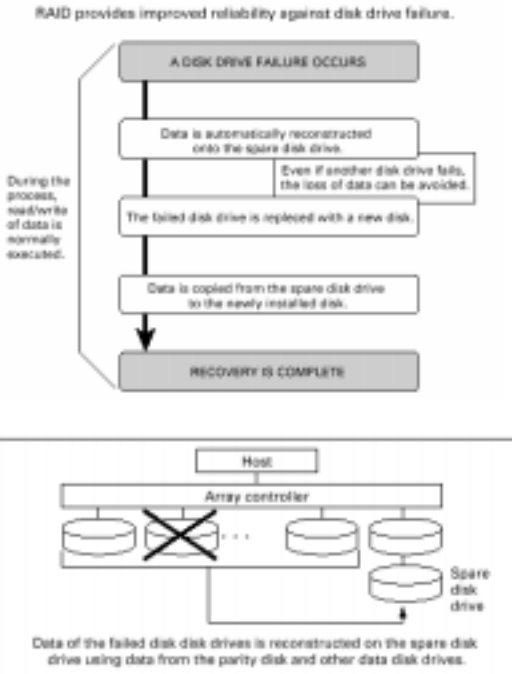
- 1) RAID Level 0: Striping—RAID Level 0 supports data striping which means data is distributed among many different disks as it is written, as opposed to sequentially writing the entire file to one disk. Level 0 is actually a non-redundant level of RAID, and is therefore somewhat of an anomaly. Level 0, however, provides the greatest level of data transfer. Because data is split across drives it can be written very quickly. Since no redundant information is stored on the disk, performance is very good, but there is no recovery from error. If an error takes place in a array configured at Level 0, then the data is simply lost. Level 0 is most commonly used in digital video applications where sustained writes in excess of 20MB per second are fairly common.

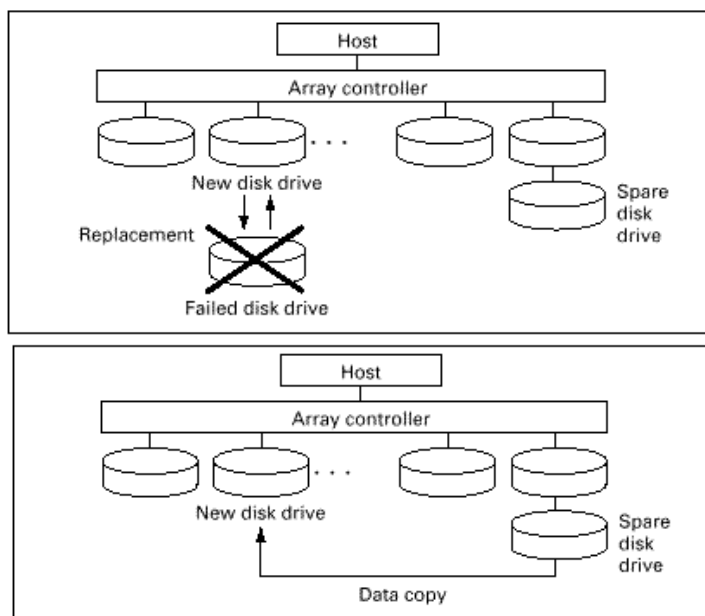
- 2) RAID Level 1: Mirroring—Level 1 provides redundancy by duplicating all data from one drive onto another drive. The performance of Level 1 is only marginally better than the performance of a single drive. In the event that either drive fails, data is still available in identical form from the redundant drive. The application of mirroring is obviously one where there is little or no time to perform backups, and data must be protected as it is being used.
- 3) RAID Level 2: Non-Error Correcting—Level 2 was originally specified for hard disk drives that did not have built-in error detection. This RAID level uses an entirely different error correction scheme called Hamming Error Code Correction. All present-day SCSI drives have built-in error correction, and since most modern RAID architectures are built around SCSI devices, this RAID level is largely dormant.
- 4) RAID Level 3: Striping & Parity—Level 3 provides both striping and parity. Data is striped across drives at the BYTE (8 bits) level. Parity is maintained on a dedicated parity drive. The speed and effectiveness of Level 3 is determined by the specialized hardware controller.
- 5) RAID Level 4: Striping & Parity—Level 4 also uses both striping and a dedicated parity drive, but data is striped at the block level. The performance of this level is designed for very active high speed reads. Large file servers and archival storage are ideally suited to this level of performance. While reads operate at speed levels close to Level 0, writes are slower because each write must also update the parity drive.
- 6) RAID Level 5: Striping & Parity—Level 5 performs striping and parity in similar ways to Level 4, but does not use a separate and dedicated parity drive to store the parity information. Instead, parity is distributed among the drives in the array. This level increases write speeds considerably since no extra cycle is required to update the parity drive. The performance for reads, however, tends to be slower than a Level 4 array. Level 5 is ideally suited for database servers, where I/O happens in small chunks and is frequently random.

As mentioned above, each RAID level provides a series of tradeoffs between I/O functionality. The fastest possible device for both is at RAID Level 0, and that includes no provisions for data redundancy. A drive failure at this level would have serious consequences.

Redundancy vs. Fault Tolerance

If redundancy is the main objective of the RAID array, then all levels of RAID except Level 0 provide it. Because of the level of data security provided by many RAID systems, the definition of redundancy has been extended to also include fault tolerance. The objective in the deployment of RAID is not to provide fully fault-tolerant operations, nor is it the primary mission of RAID to provide it. Many of the benefits of RAID can be achieved without this provision, or redundancy of data may be the extent of the fault tolerance which is required in the processing system. However, in order to view RAID as also providing some measure of fault tolerance, the protection against data loss must be extended to components of the RAID system other than the disks themselves.





Protection Against Power Loss

Many RAID arrays keep a large part of the incoming and outgoing “inventory” of data in a temporary location called a cache. Cache is typically RAM (Random Access Memory). In the event of sudden loss of power, whatever is in the cache needs to be protected so that the writing or reading operation can be continued. Should power loss occur during a write cycle the unit should be able to continue to function until that cycle is completed, or be able to remember where it was and complete the cycle when the power comes back on. Similarly in a mirrored array, if one of the “mirrors” fails, does operation continue? It probably should not, since the purpose of this feature is to provide duplication.

Component Replacement

The ability to replace a device when a failure occurs without shutting down may also be a desirable feature. For disk drive failure this is called a “hot swap”. RAID systems supporting “hot swap” will allow the broken drive to be removed while the system continues to operate. Once the drive has been replaced, the system will reconstruct the data that was originally intended for it.

Similarly, in some systems it may be desirable to replace the drive controller or other device under the same circumstances, or to replace the power supply itself while the system continues to operate.

RAID PERFORMANCE ISSUES

Out of the box, any RAID solution is likely to provide performance advantages and redundancy better than standard components. However, the storage management capabilities of certain applications software were not necessarily designed with RAID in mind. Certain databases, for example, may be “set-up” in order to provide maximum performance under different storage conditions than those provided by RAID. In some instances, the additional parity steps, in particular, may slow database performance. This is frequently a function of the database itself, and not of the RAID array. Under those circumstances, the database may need to be reconfigured and optimized for RAID before the true potential of RAID can be achieved.

Another performance issue affecting the RAID device itself, of course, is the ability to configure one device to operate on several RAID levels at once. By separating the one large RAID volume into several partitions, a truly powerful self-contained storage subsystem can be assembled. For example, one RAID volume can be configured to operate at RAID 0 to maintain the fastest possible I/O rate, while another operates at RAID 5 in order to provide speed and redundancy, and a third operates at RAID 1 in order to provide full data duplication on a second-by-second basis.

Other performance issues concern RAID product features such as bus compatibility and efficient SCSI implementation. For obvious design reasons, the “bus” carrying the data into and out of the RAID system must be able to carry the same number of data passengers as the bus carrying data passengers from the controller to the disks themselves. Otherwise some data will be stranded in a staging area. There are multiple versions of SCSI and some versions are faster than others. So it is likely that either more data can arrive than can be sent to the disks, or both are operating at lower capacities than they could be.

SCSI itself must also be effectively managed. One way of achieving SCSI management is with a concept called Tag Queuing. A number of

requests are likely to come across the SCSI cable at the same time. These requests don't always coincide with the previous activity. If one request, for example, has the drives looking for data in one location, the next request may ask them to move to another location, while the following request asks them to go back to where they were in the first place. The ability to prioritize by grouping similar requests in order to minimize frenetic efforts, provides numerous I/O advantages by allowing the system to respond more quickly.

RAID IN AN IMAGING ENVIRONMENT

As mentioned earlier, RAID offers many storage features ideally suited to an imaging environment. The most important of these are the ability to configure a single device to provide non-stop, high speed performance, and the enhanced options for data security.

Stand-alone Solutions

In order to evaluate RAID as a stand-alone imaging solution, all of the features indicated above need to be analyzed. In addition, in order to deploy RAID as a stand-alone imaging solution, or the only storage solution in an imaging environment, the following issues must be examined:

- **Document Processing Volume:** The volume of documents either coming into, or being processed on a daily basis, generally determines the kinds of performance advantages likely to be gained from magnetic storage. Cost estimates will have to be independently confirmed, but in many cases, the document volume may be high enough to make conventional magnetic storage prohibitive, and low enough to sidestep the cost effectiveness of optical disk storage. A fully-configured RAID solution, operating at three RAID levels simultaneously, may be able to provide the level of processing performance required by the system by doing everything in one box. The solution could be configured to provide high speed access, processing redundancy and data backup in a single box. Frequently

in imaging environments it is not possible to perform backups simply because the system can not be stopped or slowed down long enough to do it. Being able to mirror one set of disks may provide precisely the level of disaster avoidance required by the system.

- **Retention Period:** The length of the retention period for documents is also a critical determinant for using RAID as a stand-alone system. The cost per megabyte of storage for RAID is not competitive with optical storage. However, if documents do not have to be kept for extended periods of time, then it may never be necessary to send them to an optical device. They may be kept in RAID until the processing period is over and then either destroyed or saved to tape for long term archive. In some situations, particularly in a forms processing environment, it may not be necessary to keep the image of the form at all. During the capture process, images are digitized, the data is extracted and then the image is destroyed. RAID may provide an excellent vehicle to hold the image data during the processing cycle.

Mixed Storage Environments

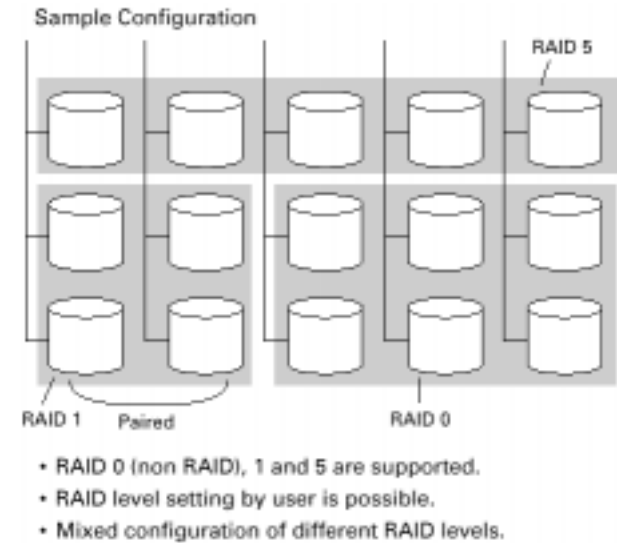
Because the typical imaging environment makes use of many storage devices, the benefits of RAID in an environment where magnetic and optical technologies are both utilized are of obvious value.

RAID is an ideal complement to the optical disk jukebox during both the jukebox writing cycle, and for pre-fetching information from the jukebox for the daily processing work. Jukeboxes have much slower data access rates than magnetic disks. They are also much slower to write information to disk. Thus, a high volume of image data coming into a system may take an extended period of time to be written to the jukebox. If the data must be available for processing during that period of time, the information must be kept on an accessible source. A RAID device can provide both high speed access to the operators requiring images, and a secure environment to hold them until the jukebox can finish its writing cycle.

Similarly, the same RAID device can be utilized as the place where images required for the next day's processing are loaded overnight from the jukebox to the RAID system. By pre-fetching images, the jukebox or other optical storage device is kept at its normal operating speed. Performance of the imaging system is greatly enhanced because the operators are primarily working with the high-speed device.

The use of storage management software, in particular hierarchical storage management software (HSM) can be an invaluable management asset to these types of mixed storage environments. By providing the ability to migrate images and data on an automatic and rule- oriented basis, the flow of images between the RAID device and the jukebox can be carefully controlled. Furthermore, HSM software also provides the facility to migrate images and data inside a multiple level RAID system. Data can be moved from the mirrored partition for backups, and data from a Level 5 partition can be placed into a Level 0 partition for extremely fast access.

Flexibility in Configuring RAID Levels



Characteristics of each RAID Level		
	Advantage	Disadvantage
RAID 0	Highest operating efficiency of disk drives	With no failure resistance
RAID 1	Reliability and availability are improved by mirroring	Low operating efficiency of disk drives because double capacity is required
RAID 5	Parity data ensures high reliability. High performance for transaction processing	None

The potential configurations are endless, and heavily dependent on the application, environment and platform. One thing, however, is certain, RAID provides a flexible storage alternative while maintaining high- speed performance and reliability, the hallmarks of an imaging solution.


AUTHOR BIOGRAPHY

The Rheinner Group is a leading research, consulting and education firm in the document imaging, management and workflow industry. Its Certified Document Imaging Architech (CDIA) Education Program, which covers many of the same issues addressed by The Rheinner Group's Technology Guides, is the most popular training program in the imaging industry. For more information on The Rheinner Group, CDIA course schedules, or to obtain help designing and implementing document imaging and workflow systems, please call 781-741-8100 or visit our web site, at www.rheinner.com.

COPYRIGHT INFORMATION

This book is the property of The Rheinner Group and is made available upon these terms and conditions. The Rheinner Group reserves all rights herein. Reproduction in whole or in part of this book is only permitted with the written consent of The Rheinner Group. This report shall be treated at all times as a proprietary document for internal use only. This book may not be duplicated in any way, except in the form of brief excerpts or quotations for the purpose of review. In addition, the information contained herein may not be duplicated in other books, databases or any other medium. Making copies of this book, or any portion for any purpose other than your own, is a violation of United States Copyright Laws. The information contained in this report is believed to be reliable but cannot be guaranteed to be complete or correct.

Copyright © 1996, 1997, 1998, 1999 The Rheinner Group



This Technology Guide is one of a series of guides, written by The Rheininger Group and distributed by techinfocenter.com, designed to put complex document management imaging, and workflow concepts into practical and understandable terms. Each guide provides objective, non-biased information to assist in the internal education, evaluation and decision making process. This Technology Guide, as well as the other Document Management, Imaging, and Workflow Technology Guides in the series, are available on

www.techinfocenter.com

[click here](#)



Distributed by

**www.
techinfocenter
.com
[click here](#)**